British soldiers of the First World War: creation of a representative sample

Lamm, Doron

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Empfohlene Zitierung / Suggested Citation:
Lamm, D. (1988). British soldiers of the First World War: creation of a representative sample. *Historical Social Research*, *13*(4), 55-98. https://doi.org/10.12759/hsr.13.1988.4.55-98

British Soldiers of the First World War: Creation of a Representative Sample

Doron Lamm *

Abstract: This article describes and evaluates the most comprehensive and systematic source known to date, of information on individuals who were British soldiers during the First World War: the Army's collection of 2 - 3 millions of soldiers service files, still colsed to the general public. The technical and methodological complexities of using these files are assessed, with special emphasis on the problem of representativeness. Two separate sections are devoted to the process of data base creation through sampling and coding. Contrary to common belief, a quantitative analysis proves the data - in their entirety - to be representative of the wartime ranks, sensitive to short time fluctuations and remarkable accurate in portraying sub-groups within the population.

1. Introduction

In 1982, while sketching the territory of Social History, Theo Barker and Michael Drake called our attention to »six million case studies«, personal service files of soldiers, »collected as a result of conscription in the Second World War... [that] await the courageous at the Public Record Office«.(1) These files, they rightly claimed, were comparable to similar data from the First World War. The authors, as well as their readers, presumably assumed that a comprehensive study of the First World War material had already been carried out. Unfortunately, this was not the case. The main bulk of millions of First World War files was never studied. The purpose of this report is to acquaint the reader with the most comprehensive historical source which now exists, on individuals who were British soldiers during the First World War.

* Address all communications to: Doron Lamm, Department of History, Birkbeck College, University of London, Malet Street, London, WC1E 7HX, England.

To be more precise, the report deals with four different historical sources. First and foremost is the archive of First World War soldiers' service files, held by the Ministry of Defence in the Public Record Office repository in Hayes, Middlesex. The nature of the archive, its structure and the actual information it holds are described in chapter two. Most of the data in the files were recorded on »Particular Instance Papers« or PIPs\ These are official forms »the subject Matter of which is the same, though each relating to a different person. ...While each individual document may be of little importance by itself, taken together...these papers enable certain broad conclusions as to historical, economic and social trends to be drawn«.(2) However, this description of an individual form undervalues our material. A dossier of several PIPs, each relating to a different aspect though all regarding the same individual, is of considerable interest even on its own. Millions of such files, taken together, make a formidable historical source.

Like most collections of PIPs the First World War archive holds a massive bulk of material. It is impractical for historical study to use all the documents, which must undergo a process of selection. But creating a representative sample is rarely a smooth process. Chapter three is devoted to a detailed account of the sampling process. It concludes with a brief description of the second historical source: a representative sample of the First World War archive, also held, though separately, in the Hayes repository.

The last two sources appear in the form of machine readable databases. The Recorded Database contains only a fraction of the information which is available in the files. This information is now stored on a computer, readily available for analysis. It is described in chapter four. Chapter five gives a detailed account on the transformation of the recorded data into a numerically coded database. It explains the reasons for coding and its repercussions on the data, lists the codes of most of the variables, and discusses the guiding assumptions for the geographical and occupational codes. As chapter five explains, coding is interpreting. Therefore, of all outsources the coded database is the only one that is likely to be altered.

The rest of the report is devoted to an assessment of the representative sample and the archive alike. This is necessary because of the notoriety and suspicion which accompany the First World War archive. The poor condition of the documents and the confusion surrounding their history since 1920 have led archivists to believe that: a. Due to loss and decay of files the archive is hopelessly unrepresentative of the army. b. The information in the available files is practically inaccessible, and c. Of all the archive, the only set of documents which is worth saving is that called the »1914-1920 Collation«.

There are two complementary approaches to the problem of representativeness. The first is to enquire into the history of the files, especially the missing ones. Here one poses questions such as »why have certain files never been sent to the archive?« or »why had certain files been sent out?« »what made bureaucrats return some files but not others?« etc. There are several problems in pursuing this strategy. It is likely that there will be not one general answer to these questions but rather several different answers to each query. There is also good reason to believe that not many answers will be forthcoming, for the process of archival accumulation was both undocumented and dispersed. Finally, even if we get some insight into the process of selection, there is still no guarantee that our newly-acquired knowledge will throw much light on the initial problem of representativeness.

The other approach is a comparative examination of the existing material. This line of investigation requires a trustworthy set of published statistics on the wartime army, to serve as an archetype of the complete archive. A sample that will be drawn from the archive could then be compared with this archetypal population and reveal the goodness of their fit. Here one risks using defective statistics which may impose a biased model upon the sample and reduce its analysis to a mere reflection of orthodoxy. In addition, one should satisfy oneself that the variables compared are relevant to the question of representativeness. In spite of these dangers, the second strategy - cautiously pursued - holds more chances for success, and it also has the advantage of tackling the questions of data accessibility and future preservation of the documents.

Chapter 6.1 proceeds, therefore, to consider the adequacy of the sample as a proxy for the First World War British ranks. Putting it crudely, our assumption is that the preservation of the files was random, and that our representative sample was drawn from a huge random sample of the army. In the absence of adequate information on the army 1 sometimes resort to comparisons with the civilian population. There, the assumption of representativeness can not be valid. One should not expect the army to possess the same attributes as those of the whole population. A certain inherent variance between army and society must be allowed for, and a more general agreement is then expected.

Chapter 6.2 surveys the difference between the two sections of the archive, and establishes, in broad terms, their relative contribution towards its representativeness. Consequently, it also dwells briefly on future preservation policy.

This report is aimed at both researchers who might wish to use any of the aforementioned sources and at the civil servants entrusted with the task of preserving the original documents. Consequently, it touches upon several different issues. In part it is unavoidably detailed and technical. I

have also tried to ensure a coherent narrative for those who are only interested in some of these issues. Readers interested in the raw data can avoid chapters four and five on recording and coding. For those who are involved in the future of the archive, chapters three and six may be of special interest.

2. The First World War Archive of Soldiers' Service Files.

2.1 The Men.

Patterns of recruiting to the British army changed dramatically in August 1914. For generations, recruiting officers were subject to the unsympathetic mechanisms of the labour market, where they had to compete with civilian employers. They relied on economic depressions and seasonal ebbs in the demand for labour to swell the permanent reservoir of the unemployed, and dreaded economic tides which contracted it through opportunity and higher wages. For it was mainly young unemployed, unskilled labourers and the »unemployable residuum« who, at one time or another, preferred a soldier's career to the hardship and occupational insecurity of civil life.

The army itself was small in comparison to other continental powers. Throughout the nineteenth century its peacetime establishment rarely exceeded 200,000 men. Indeed, it was small enough to be able to pick and choose its recruits even when the candidates' pool contracted. Facing the Treasury's reluctance to increase spending on soldiers' pay, the army usually responded to scarcity by lowering the physical requirements of new recruits. Recent estimates have shown that a higher proportion of working class men than was customarily perceived were at one time or another candidates for recruiting.(3) Those who were finally approved for service should not be regarded as the scourings of the barrel, though they often were by most of their contemporaries. Nevertheless, in times of great need such as the Boer War, when the army's annual recruiting figures for 1900 peaked with 98,381 new men, medical standards for recruits were so low as to suggest that the traditional recruiting potential was fully exhausted.(4)

The rush to the colours during the early months of the First World War radically widened and diversified this long established constituency for recruiting. Often, during September 1914, a year's vintage of peacetime recruiting poured upon the helpless recruiting officers in a single day. 600,000 men volunteered in that month alone.(5) By Christmas 1914 almost 1.2 million men enlisted. These NCO's and men came from all parts of the country and walks of life. The various sectors of the economy, age

groups and social strata were now represented in the army as they had never been before. The Victorian and Edwardian professional army turned overnight into a »Nation In Arms«. Four years later, on the eve of the armistice, long after the introduction of conscription and dilution, more than 6 millions, nearly a half of the prewar, occupied male workforce, had enlisted.

2.2 The Files and the Collations.

On enlistment, each of these candidates for recruiting had to submit certain particulars regarding his civilian life, occupation and family. The information was recorded on his Attestation Form. Then he was measured and medically examined by a medical officer who filled in his Medical History form and declared him to be either »fit« or »unfit« for service. These two forms became the cornerstone of each individual's service file. As the war went on each file accumulated more information as well as additional army forms.

By the end of the war these millions of files comprised an unprecedented compilation of data not only on the wartime army but also on Late-Victorian and Edwardian society. Their significance did not escape the notice of contemporaries, nor indeed of later historians and social researchers. Already in 1919 proponents of eugenic racial amelioration as well as their adversaries sought to exploit the information to fuel their long fought battles over social policy. Studies of prewar health standards tended to make use of aggregated statistics derived from some of these files by the Ministry of National Service Report of 1919.(6) A major deficiency of the data - the omission of women and non-combatant members of society - was usually ignored. Children, women, the aged and vulnerable groups in society are, the most significant groups for research on human welfare, though the least documented ones. In this respect the First World War files are no exception. Yet, it would be equally wrong to overlook the available information on these groups, which accompanies the data on male soldiers. Parents, wives, children and other siblings do feature extensively in the documents, and shed light on topics which »soldiers only« data cannot.

Until recently interest at the micro level of the individual file was confined to many thousands of siblings and genealogists. The spread of Social Science History and its methods of research have brought such documents to the forefront of historical research. It is the availability of mass quantities of standardized data at the micro level which appeals to social and economic historians, historians of modern armies, demographers and anthropologists. Nevertheless, these files have never been the subject of academic research. One obvious reason for that is purely administrative. In

view of the private nature of the information recorded in the service files, they were put under a 75 years closure, to be opened to the general public only in the mid 1990's.

Unfortunately, not all these files have survived. To date, we have at our disposal, at the Public Record Office's repository in Hayes, Middlesex, between two and three million files of NCO's and men who were discharged from service between 1914 and 1920. They are kept in two separate, alphabetically arranged collations: »The Burnt Documents Collation« and »The 1914-1920 Collation«.

Information regarding the history of the files since 1920 is depressingly patchy and incoherent. Initially, they were kept in regimental and local record offices. In some isolated cases, regiments such as the Black Watch have retained their files until this very day. More commonly, however, files were sent to a Ministry of Defence (MOD) central repository some time after becoming inactive. There they were stored according to the soldiers' period of discharge, which in this case encompassed the years 1914-1920. The sub-division of the archive at that stage was probably according to corps and regiments. Many other files were sent by local and regimental record officers, as well as the central repository itself, to various departments of state, notably to the Ministry of Pensions. A large proportion of these files was destroyed as recently as 1980 (!).(7) During the Second World War the MOD's central repository of First World War files was hit by a bomb. An unknown number of files was completely destroyed. The remainder, which still forms the greater part of the present archive, was damaged by direct fire, heat and eventually, water. Some time after the Second World War this remainder was rearranged in alphabetical order and received the title »The Burnt Documents collation« ('BD'). Meanwhile, files which kept on arriving at the MOD's repository, were collated separately, and now comprise the »1914-1920 collation»('14-20').

The physical condition of the First World War files in Hayes varies considerably. The obvious distinction between burnt and non-burnt files is insufficient and could at times even be misleading. It is important to note that the 'BD' do not contain only burnt or damaged documents. For example, the 'BD' also holds a considerable number of 'Z' files, distinguishable by their paper overcoats marked with the letter 'Z', meaning transferred to army reserve on demobilizations The 'Z' files did not form a part of the collation at the time when it was damaged. Files of soldiers discharged prior to 1914 or post 1920 can also occasionally be traced in the 'BD' as well as other non-burnt war time files which found their way to the collation for no apparent reason. Although this report does not provide an exact estimate of the proportion of undamaged files in the 'BD', my impression is that it is by no means negligible.

The very term »Burnt Documents« is also a poor indicator of the accessibility of the information in the files. There is a great variation in the levels of damage within the collation. The damaged files in the 'BD' seem to have suffered mostly from excessive heat and water. The heat consumed the outer margins of the paper, progressing towards the centre of each document. While some of them were reduced to rather small - at times, insignificant - remnants, most of the existing documents only lost their edges, leaving their information legible. Still, these documents require careful handling. Many »burnt« leaves of paper lost their flexibility and turned fragile. Those of the files which were affected by water and subsequent dampness can be even more difficult to use; here, in the most severe cases the water washed the ink off the paper, leaving only faint marks which are barely legible. It also made the papers stick together, so that separating them often entails loss of documents. But all these problems do not apply to all the files. And on the whole, the Burnt Documents still comprise a manageable, legible and uniquely rich source for historical research on individuals and their environment prior to and during the Great War.

In some respects the files in the 'BD' are superior to the '14-20' documents, a great number of which have been weeded and standardized. Most of the '14-20' still contain principal forms such as Attestation Form, Medical History sheet or Casualty Form. They do not contain, however, material which is abundant in the 'BD' files such as correspondence of the soldier and his family with army officials or civilian documents such as marriage certificates or birth/death certificates of children. Nor are they likely to contain employers' letters of reference, odd reports or particular instance forms regarding, for example, repatriation, wills or a dead soldier's living siblings. On the other hand the '14-20' files contain more information regarding injuries and disabilities, reports from medical boards and subsequent army pensions. The emphasis of the '14-20' files on the soldier's medical condition at the time of his discharge occasionally comes at the expense of information regarding his situation on enlistment or particulars on his family. For instance, a duplicate of a Medical History sheet, which was prepared only for the use of a medical board investigating an injury, tends to omit »irrelevant« information such as place of birth, or occupation. Indeed, a comparison of the availability of variables in each of the collations revealed that apart from variables concerning the discharge from service, all other variables were more available in the 'BD'.

A more fundamental difference between the collations is revealed when variables related to individuals' histories of service are compared. Such a comparison strongly suggests that each collation represents a different composition of soldiers as far as their wartime experience is concerned. Since wartime experience is deemed to be all but independent of non-

military factors such as age, civil occupation, social status or physical fitness, there is a good reason to believe that the social and demographic character of each collation may also be different. This phenomenon appeared to be most striking in the comparison of the reasons for discharge of some 200 soldiers sampled randomly from both collations. See table 2.1:

Table 2.1: Reasons for Discharge from Military Service. (%)

	'BD'	'14-20'
Discharged on demobilization	65	14
Killed or died while in service	14	2
Discharged on medical grounds during the war	11	43
Discharged as »unlikely to become an effective soldier«	1	21
Discharged on administrative grounds	6	19
Deserted	2	0
total	99%	99%

Source: D. Lamm, A Report on a Pilot Sample Taken from First World War Data at the Ministry of Defence's Archive at Hayes.

While two thirds of the 'BD"s soldiers served with the colours until demobilization, a similar proportion of the '14-20' was discharged as unfit before the end of the war, mainly on medical grounds. Many of them were war casualties, but the health of many others had already been impaired prior to enlistment. This was especially apparent among those who were regarded as »unlikely to become effectives This is not too surprising considering that most of the files in the '14-20' arrived at Hayes after being at the disposal of the Ministry of Pensions and subsequently the Department of Health and Social Services. A further comparison of military histories of soldiers in both collations is carried out in chapter 6.2 below. Here 1 am merely seeking to point out that the differences between and within the collations manifest themselves not only in terms of form but also in contents.

2.3 The Forms.

Differences in the fate of soldiers are only one obvious reason for a wide variation in the sort and quantity of army forms in the archives' files. As I have shown, even the history of the files themselves had some lasting repercussions on their content. Certain branches of the army used different varieties of forms on similar occasions such as attestation. These documents differed in form and contents. Forms also tended to alter with

time in order to suit the changing needs and emphases of the army's bureaucracy. It would be impractical to describe all these different forms in detail. Instead I shall focus on six forms which are widely available and provide relatively more information than others.

Attestation form(s): [A.F. B250; B2512; B2085;]

This provides names, army number(s), units and ranks, dates and places of attestation, joining and approving of the soldier as well as names and ranks of the NCO's and officers in charge of these stages of the recruitment process. It also gives information on age, place of birth and/or address (parish/town/county), nationality, occupation, marital status, religion, residential status, apprenticeship, prison record, previous military experience (incl. rejections), physical measurements (height, weight, chest measurements) and visual description on attestation. Also included are names and addresses of siblings (»next of kin«) and in case of a married soldier or a father, the places and dates of marriage and births of children. The Attestation Form was regularly updated during service to provide information on ·»transfers promotions reductions casualties etc.«, wounds, participation in campaigns, subsequent decorations, and a summation of periods served »abroad«. The regular updating also regarded changes in marital status, in family composition (births & deaths) and in the next of kin's address. Finally, it provides the dates and administrative reasons for discharge.

By combining together this wealth of detailed and varied information the Attestation Form constitutes the single most important item in each file. Since it was meant to provide the outer cover for the rest of a file's forms, it was printed on a thicker, superior type of paper, and enjoyed a higher degree of preservation.

Medical history: [A.F. B178; B178a;]

The Medical History sheet is divided into four tables.

Table I, »The General Table«, provides names, corps, places and dates of enlistment and medical examination, place of birth (parish, county), age and occupation. It then lists physical measurements (height, weight, chest measurements), assesses physical development, records vaccinations and vision and specifies congenital peculiarities and »slight defects not sufficient to cause rejections Medical grades were included after their introduction in 1916. The table concludes with the date and reason for the soldier becoming ineffective.

Tables II - IV were designed for current use during service. Table II recorded admissions, discharges and medical treatment in hospitals. Table III gave dates and a brief reference to medical boards, courts of enquiry, recent vaccinations, dental treatment etc. Table IV was designed to record transfers, arrivals and departures from military stations. It was rarely used.

Casualty Form - Active Service: [A.F. B103; B103-2]

This form was designed to »record promotions, reductions, transfers, casualties etc.« and to specify the source from which the information was received. It also includes entries for religion, age on enlistment, date of enlistment, date from which service is reckoned, military qualifications and civil occupation. In late 1917, initials regarding marital status (M,S) and occupational coding were added to all casualty forms. This form also gave the most up to date address of next of kin. It was regularly updated and usually gave the most comprehensive description of service

Statement as to Disability: [A.F. Z.22]

This form was filled in by all soldiers who went through the demobilization process. The first section of the form recorded names, military unit, regimental number etc. It also included age, date, place of recruitment and medical category when first joined, cause for discharge and prospective address after discharge. Only in a case where a soldier was claiming impairment in his health due to military service he was asked to fill the following details: Where did he serve? What was his medical complaint and what might have caused it? At which place and date did it originate? The soldier was then asked to specify the military hospitals where he received treatment, and to provide the name of his National Health Approved Society, as well as the names of civilian doctors who used to treat him. The concluding section of this part referred to his civilian occupation, the name and address of his last employer and the capacity in which he was employed prior to enlistment. This information was followed by the opinion of a medical officer who was asked to provide diagnosis, to state whether any disability was caused or aggravated by military service, and to assess the degree of disablement.

Medical Report on a Soldier Boarded Prior to Discharge or Transfer ... to the Reserve and Medical Report on an Invalid: [A.F. B179a; B179]

This form(s) includes a »statement of case« made by a medical officer which was then brought before a medical board. The board reexamined the soldier and summarized its findings in the »opinion of the medical board«. If invalidated, the form was used by the Ministry of Pensions for determining the level of pension. The statement of case recorded the date of origin and nature of the disability. It gave a concise history of medical treatment and an assessment of the present condition. It concluded with a recommendation for discharge or transfer. The »opinion of the medical board« considered the origin, nature and the present condition of any disability claimed or discovered. It specified the degree of disablement at present (in %) and recommended as to future service. These summaries and comments were usually laconic, though they varied in length and detail.

Regimental Conduct Sheet:

This form briefly described those offences committed by the soldier which evoked disciplinary procedures against him. It stated the nature of each offence, distinguished between drunkenness and all other offences, and specified the date and place where it had been committed. The names of the judge(s) and witnesses came alongside a short summary of the verdict and »general remarks«.

Three major apparent deficiencies of the archive are, therefore, the obvious overlap among forms, the great diversity of aspects represented in each file, and the huge volume of the data. It is worthwhile to point out, however, that these alleged weaknesses could be regarded, in fact, as contributing to the archive's quality as a historical source. Overlapping data, for example, are very often an advantage; in most files some of the forms are missing and others are incomplete. Thus, the more times a variable was recorded, the higher the chances that it would be available. Moreover, re-occurrence of variables helps to detect inaccurate information and frequently is the only means of correcting it. Likewise, partial or ambiguous entries, such as »a foreman« for occupation, are often completed by another document which states »a foreman in boot & shoe factory«. This ability to corroborate and clarify self-declared information is a major contribution to the archive's credibility. It is, therefore, important to realize that no weeding out of documents - as shown in the case of the '14-20' - will be harmless to the archive. While researchers are justified in collecting only the information that is relevant to their research, a permanent exclusion of data as »unimportant« is likely to diminish, or even destroy, the archive's significance as a historical source.

3. The Creation of a Representative Sample

The purpose of sampling from any archive is to produce its mirror image on a much reduced scale, which although smaller in numbers is nearly as significant a set of data for research as the original archive itself. The obvious advantage of a sample is that it is quicker to record than a full archive and is easier to manage. However, it is only when all the different sections of the archive are represented in the sample in sufficient numbers and according to their relative share, that the sample can be regarded as a scientifically accurate substitute.

3.1 Sampling the Archive.

In the case of the Hayes archive such an equidistributed sample could be drawn either randomly or systematically. Recall, however, the special peculiarities of the Hayes archive. It consists of two collations, each containing an unknown number of files, arranged in an alphabetical order and stored in boxes and bundles of various sizes. Under these circumstances neither method of sampling is free of complications. A systematic sample - where, for example, each 10th box is picked - is based on the assumption that all boxes contain the same number of files, or, at least, that boxes of all sizes are spread throughout the archive with no underlying pattern. Such an assumption is totally inappropriate in the case of the Hayes archive, where different segments of each collation are characterized by either small, medium or large boxes. The absence of current enumeration in each collation, of either files or boxes, is a further complication. In order to sample systematically, we need to divide the total number of boxes by the number of required observations, the result being the interval between each observation. In the absence of such a total, the procedure for obtaining the sampling interval is cumbersome. The absence of enumeration is even more hindering in the case of a random sample where a list of »random numbers« has to correspond to an existing enumeration covering the entire population. A random number should identify one, and only one, observation, be it either a file or a box.

For these reasons it proved beneficial first to define the relative share of each collation in the sample and only then to devise equidistributed sampling procedures for each collation. Since over-representation of one collation is relatively easy to correct either by applying weights to each collation during analysis, or by discarding surplus data altogether, a higher priority was given during sampling to ensuring accurate representation within each collation.

The most reliable information regarding the relative size of each collations are the MOD's official figures on allocation of shelving space between them. The 'BD' holds 13,528 ft. of shelving space or 738 of the whole of the First World War archive. The '14-20' occupies only 5083 ft. or 27% of the total. Accordingly, the intended sample of 6700 observations would be comprised of 4890 and 1810 observations drawn from each of the collations respectively. Note, however, that by adopting this estimate it is assumed that the average number of files per shelf is equal in both collations.

3.2 Sampling the 'Burnt Documents',

A close examination of the boxes in the 'BD' revealed past attempts to assign current numbers to various parts of the collation. Fortunately, it proved possible to make use of these fragmentary numerical series in order to construct a five digit system of enumeration, which covered the whole of the 'BD\ Columns 3 and 5 of table 3.1 give the original numbers on the boxes and the corresponding five digit enumeration. By assigning numbers of the range 00001 to 99999 to the total of 32,629 boxes in the 'BD' (see bottom of column 4), only 32.6% of this range became occupied. Thus, in order to compile a list of 4890 five digit random numbers, a list of 15,000 five digit random numbers was needed. All random number that did not correspond to numbers on boxes were deleted from the list, leaving a total of 4893 relevant numbers.

It is important to note that the assignment of a single random number to one, and only one, box, was equivalent to assuming that all the boxes in the 'BD' contained the same number of files. Since the 'BD' contained small, medium and large boxes (see column 2, Table 2.1) this assumption had to be put right. In view of the relatively small number of medium sized boxes, it was possible to consider them as small boxes, without introducing a significant bias into the sample. Thus, a single file was sampled from small and medium boxes while two files were drawn from a large box. This procedure offsets the bias which might have occurred in the sample by a »one number to one box' ratio.

The single file that was taken from a small or a medium sized box was usually the fifth file from the top. In those cases where the fifth file did not satisfy a minimum standard e.g. lacked information on height or occupation, either the fourth or the sixth file was chosen. More than one file was sampled from a small box if its number appeared more than once on the random numbers list; these files were sampled in intervals of five: the fifth, tenth etc. The additional file to be sampled from a large box was picked at random from nearer to the bottom of the box. No more than two files were sampled from a large box. In order to facilitate future retrieval of sampled files, each box which contributed a file(s) was marked accordingly on its outside.

Table 3.1: The Structure and Enumeration of the »Burnt Documents« Collation.

(1) From Name	to Name	(2) Box Size	(3) Original Numbers	(4) No. of Boxes	(5) Relevant Range for Random Numbers	(6) No. of Relevant Random Numbers
Aageson T	Earl T W	S	1-10,000	10000	10,001-20,000	1516
Earle T W	Kershew V	S	1-10,000	10000	00,001-10,000	1501
Kershew Wm	Mezzett J	S/M	1-3923	3923	20,001-23,923	577
Miall A	Ozzard W	S/M/L	1064-2112	1049	31,064-32,112	134
Pablo C E	Pyzer T	M/L	1-905	905	40,001-40,905	149
Qua W	Qwy W	S	5523-5580	58	25,523-25,580	11
Raban H	Rynes G H	L/M	1-911	911	50,001-50,911	147
Sabin C R	Szvelins V	L	1-1631	1631	60,001-61,631	243
Taafe A G	Tompkins W	S	7967-9263	1296	27,967-29,263	193
Tompestt A	Tyzzer H	L	1297-1638	282	71,297-71,638	50
Ubank J W	Walker W	S	9473-10000	528	29,473-30,000	65
Walker W	Wyville W	L/M	3001-4800	1800	83,001-84,840	274
Xavier	Zyozyski	S	4719-4964	246	94,719-94,964	33
Total				32629		4893

3.3 Sampling the '1914-1920' Collation.

The '14-20' documents are kept in medium sized boxes and in bundles of various sizes. They are divided as follows:

Table 3.2: The Structure of the 1914-1919 Collation.

Letters	Stored in:	Enumeration	Standardized No. of Boxes
0)	(2)	(3)	(4)
A - M	boxes	yes	7400
N - Sx	bundles	no	1694
Sx Z	boxes/bundles	no	1177
total			10271

Unlike letters N-Z, letters A-M are ideal for random sampling. They are stored in boxes of identical medium size, seven boxes per shelf, and are enumerated from 1 to 7400. The multiplicity of sizes among the bundles and the absence of bundles' enumeration rule out random sampling in the latter part of the collation. Letters Sx Z complicate sampling even further, as they occupy only a part of their shelving space, i.e. boxes and bundles are scattered along partly empty bays of shelves.

Sampling from the '14-20' was therefore partly random and partly systematic, keeping a constant ratio of file per »medium sized boxes« throughout the collation.

Letters N-Sx occupied a net of 242 shelves. Although the files were kept in an unknown number of bundles, their shelving space was equivalent to one which could occupy 1694 medium sized boxes. The same calculation was applied to the SxZ bundles. The rest of the unnumbered boxes was simply counted. Totals are summarized in table 3.2, column 4. In order to compensate for the empty shelving space in the latter part of the '14-20',I reduced the total number of files to be sampled from this collation from 1810 to 1700.

Letters A-M were sampled according to a list of 1225 random numbers. This list was compiled by the same method used Tor the 'BD'. The rest of the collation was sampled systematically according to the ratio derived from dividing 7400 (A-M boxes) by 1225 (random numbers). A single file was therefore drawn from each sixth box. Since each shelf of bundles was regarded as holding seven boxes, each shelf contributed one file, and two files were sampled from every sixth shelf. One file was also drawn from each sixth consecutive unnumbered box. As in the 'BD', boxes and bundles from which files were drawn were marked accordingly on their outside.

3.4 Conclusions.

Sampling according to the above procedures evenly spread the choice of files within each collation as well as within the archive as a whole. Each of the newly created sets of documents is, therefore, representative of its respective collation, and together they accurately represent the full archive.

This newly created, third, collation is currently kept apart from the original archive. It consists of some 7000 files, divided into two parts:.

1. Boxes 1-105 contain the files drawn from the 'BD'.
2. Boxes A01-A27 contain the files drawn from the '14-20'.

There are considerable advantages in keeping this sampled collation separate from the original archive. First and foremost, the reasons concern the future use of the archive's documents. The First World War archive in Hayes is of huge dimensions; it is alphabetically arranged on a national level, allegedly unrepresentative and notorious for the poor quality of its documents. At present, therefore, any non-genealogical research seeking to make use of the archive must be preceded by an extremely time-consuming and costly process of sampling. However, the existence of such a representative sample renders this task unnecessary. Future researchers can use the whole of the sampled collation, or parts of it, for their own ends. This also applies to investigations carried out by the Ministry of Defence or the Public Record Office concerning the fate of the archive. Here, the sample could be used to assess the physical conditions of the documents in various parts of the archive, or to administer pilot schemes in search of new methods of preservation. Genealogical research would not suffer from keeping the sample detached from the archive. The marking of the boxes from which files were drawn should alert the paper keepers to the possibility that a missing file may be kept in the sampled collation. By a brief reference to a nominal list of all the sampled observations that was handed to them, they could locate it in the separate collation if it is indeed there.

4. The Recorded Database.

The choice of the variables to be recorded was designed to support a quantitative study of living standards among civilians prior to their enlistment. Information that was regarded as irrelevant for this end, non systematic, too impressionistic or exceptionally difficult to standardize was, in most cases, left out.

The main bulk of the data was recorded with the aid of a computer and a prompt program which served as a standard questionnaire. The program

prompted a query on the computer's screen e.g. »Town of birth?«, saved the typed answer and prompted the following query. A full questionnaire of thirty-two queries was applied to each of the sampled files.(8)

In order to minimize loss of information, to retain flexibility and to facilitate future use of the database by researchers other than myself, I did not apply numerical codes to any of the variables while recording. Occupations or placenames, for example, were copied from the forms word by word. However, for those variables where only a small range of categories was applicable, as in the case of religious denominations, I used a list of alphabetical codes - abbreviations or initials - instead of the original information. Thus, COE always stood for the Church of England as WES stood for WESleyen. Table 4.1 lists the recorded variables and provides some additional information on the data and the alphabetical codes.

In addition to the database created by the questionnaires, three other sets of data were recorded. The first was concerned with the changes that took place in the marital status and families of married soldiers prior to and during their military service. It also includes single recruits who got married while in service. These data provide the soldier's date of marriage, the dates of birth of all his children who were under sixteen years of age on his recruitment, and those born between enlistment and discharge. In case of a child's death, the precise age on death was recorded, enabling differentiation between neonatal and post natal mortality. Where reported, the specific causes of death were also recorded. Since this information affected the amount of separation allowances paid to the soldier's family, the army insisted on it being supported by official documents.

The second set of data is solely devoted to information extracted -from Army Forms 5080 on a dead soldier's living siblings. These forms specify the gender, age and place of residence of a dead soldier's widow, children, parents, brothers and sisters. They therefore provide unique information on the age structure and spatial spread of those families at a specific point in time, usually 1919. The forms were completed by a member of the dead soldier's family and approved by a minister of religion or a justice of peace.

The third set of data provides additional details to the broad and administrative classification of the »Reasons for Discharges (see Table 4.1, note 7). This is predominantly medical information extracted from medical boards reports. It specifies for example, whether a soldier was found »No Longer Physically Fit« because of »wounds received in action« or »inguinal hernia aggravated by military service«.

Table 4.1: A Standard Questionnaire for a Soldier's File.

1. Current Number (1)
2. Box Number
3. Last name and initials
4. Parish of Birth
5. Town of Birth
6. County of Birth
7. Age on Last Birthday (2)
8. Occupation
9. Marital Status (3)
10. Month of Attestation
11. Year of Attestation
12. Place of Attestation
13. Month of Recruitment
14. Year of Recruitment
15. Place of Recruitment (4)
16. Corps (5)
17. Month of Discharge (6)
18. Year of Discharge
19. Reason for Discharge (7)
20. Parish of residence
21. Town of Residence
22. County of Residence
23. Next of Kin (8)
24. Address of Next of Kin
25. Number of Children (9)
26. Served Abroad (10)
27. Religious Denomination (11)
28. Height (12)
29. Weight (13)
30. Width of Chest in Full Expansion (14)
31. Expansion of Chest
32. Further Details (15)

Notes for Table 4.1:

1. The numbers in variables # 1 and #2 refer to the enumeration within the separate collation of sampled files.
2. Information recorded on recruiting. Particulars of recruits were also recorded on attestation, prior to recruitment. The time lag between the two dates widened considerably during the »Derby scheme« and after the introduction of conscription. The information recorded on recruitment was the most accurate and consistent with other variables. This remark also applies to variables #'s 8-9, 20-24, 28-31.

3. One character alphabetical code: Mmarried, Ssingle, W-widower. Cases of separation or cohabitation without marriage were recorded as »M« and noted in »Further Details'.
4. Usually indicated the site of the soldier's unit headquarters, and had only a faint association to the soldier's place of residence. It was recorded only in the absence of information on »Place of Attestation'.
5. Abbreviated, e.g. RE (for Royal Engineers), RAMC, INF, etc. »Corps« denotes only the unit which the recruit initially joined (though not the Training Corps). Subsequent transfers to other units were ignored.
6. In the case of demobilized soldiers the »Date of Discharge' refers to the date of leaving the dispersal centre. The formal transfer to the reserve took place approximately one month later.
7. A 3-4 characters alphabetical code. The following categories for the »Reasons for Discharge« refer to specific sections of the Army Orders, para 392 of the King Regulations: DEM - demobilized, see A.0.392 xv; END - the termination of his period of engagement, see A.0.392 xviii,xxvi,xixxxii; IMPR having been convicted/imprisoned/sentenced to penal servitude by civil power, see A.0.392 x,xii; MISC - misconduct, see A.0.392 xi,xiii,vii,viii; NLPF - no longer physically fit for war service, see A.0.392 xvi; NOTL - not likely to become an effective soldier, see A.0.392 iii; NREQ · his services being no longer required, see A.0.392 xxv,ix; OAGE - old age, see A.0.392 xxiv; PUR - having claimed it on payment, A.0.392 v,xiv; UAGE -having made mis-statement as to age, A.0.392 vi; WORK - free to take civil employment which can not be held open, A.0.392 xv.
8. One Character alphabetical variable. Its various categories and their corresponding codes are listed on page 31.
9. Number of living children who were under sixteen years of age prior to the soldier's enlistment, and children who were born between recruitment and discharge.
10. One character alphabetical code: N - served in Great Britain and Ireland; Y · was despatched abroad.
11. 1-3 characters alphabetical code. The various categories and their numerical codes are listed on pages 29-30.
12. Expressed in inches and tenth of an inch.
13. Expressed in pounds.
14. Expressed in inches; also applies to variable # 31.
15. Used for consistent recording of the following variables:a.Living Out of Father?; b.Apprenticed?; c.lmprisoned prior to service?; d.Previous military service?; e.Previous rejection from service?; As well as date of marriage, dates of birth and deaths of children and reasons for deaths of soldier and children.

5. The Coded Database.

Quantitative analysis of a database of such a size and nature requires certain procedures of standardization. Since variables were fed into the computer in a fixed sequence, it was possible to type-in identical entries which, in fact, had different meanings. For example, »Essex« as a reply to the sixth query of the questionnaire would mean »county of birth: Essex«, though as a reply to the twenty-fourth query should be read as: »address of next of kin: Essex«. The data had, therefore, to be arranged in a »fixed format« where the position or »field« allocated to each variable, gives the entries their exact meaning. Moreover, since the data were entered in their original form, they featured all sorts of spelling variations and abbreviations. »Beds«, »Bedfords.« and »Bedfordshire« are typical examples. No statistical package, however, can cope with such variations in the data unless they are first coded. Only then can the computer recognize them as being the same category.

Coding, however, is primarily a process of inference and interpretation. Obviously, there is not much scope for interpretation while coding simple variables such as »Marital Status' or »Next of Kin'; here, the small range of applicable categories makes them easy to manage. It is more difficult to code places: even if a discrete code is allocated to every category, one would still need to solve, for example, the problem of several places bearing the same name. Moreover, there are alternative underlying concepts for coding places into groups, the choice of which depends on the aims and methods of the research. Administrative breakdown into counties, a distinction between rural and urban districts or a distribution according to levels of income are just three legitimate categorizations. Occupations pose an even greater problem. The meaning of an occupational title is determined by the context in which it is examined. The title »a solicitor's clerk«, for example, changes its significance over time and contains a variety of implications as to social status, level of income, training and education, job security and so on. If codes are to be comparable as well as manageable, they should consistently represent only some of these meanings.

It is obvious, therefore, that no coded information can fully replace the original data. Technically, a coding system merely creates a new, separate, set of data, leaving the original intact. The original sources - the actual files or the recorded database can then be used for clarifying ambiguous entries or for adding uncoded information to the analysis. Most importantly, it will retain its potential to be re-coded according to different concepts or for different purposes.

In this study, while running the coding program, the computer uses directories of specific variables which combine alphabetical categories

with their numerical codes. When it equates the alphabetical entry »Lancashire« with the alphabetical part of the category »216Lancashire« it produces a new database where the word »Lancashire« is replaced by the code 216. An entry that is missing from its respective directory would not be coded. Thus, when new observations are added to the recorded database, the directories would usually also be updated, in order to include any missing categories.

5.1 The Coded Variables.

Out of eighteen alphabetical variables in the standard questionnaire, (table 4.1) fifteen variables were coded. To these should be added the five variables recorded by the »Further Details« query. (Table 4.1, note 15) The coded variables are listed in table 5.1 below. The codes for all variables, excluding places and occupations, are given thereafter. A short description of the principals underlying the coding of places and occupations concludes this chapter.

Table 5.1: Coded Variables.

Variable Name	No. of Digits in Code
1. Parish of birth	3 digit(s)
2. Town of birth	3 "
3. County of birth	3 "
4. Occupation	10 "
5. Marital status	1 "
8. Place of attestation	3 "
7. Place of recruitment	3 "
8. Reason for discharge	3 "
9. Parish of residence	3 "
10. Town of residence	3 "
11. County of residence	3 "
12. Next of kin	3 "
13. Address of next of kin	3 "
14. Served abroad	1 "
15. Religious denomination	2 "
16. Residential status	1 "
17. Apprenticeship	1 "
18. Imprisonment	1 "
19. Previous military service	1 "
20. Previous rejection from service	1 "

5.1.1 One digit codes:

a. Variable name: MARITAL STATUS
Codes: »1« Single
»2« Married
»3« Widower
» « Not available

b. Variable names: SERVED ABROAD;
LIVING OUT OF FATHER'.'
APPRENTICED?
IMPRISONED?
PREVIOUS MILITARY SERVICE?
PREVIOUS REJECTION?
Codes: »1« Yes
» « No/Not available

5.1.2 Two digit codes:

a. Variable name: RELIGIOUS DENOMINATION
Codes: »10« Church of England (COB)
»20« Presbyterian (PRS)
»30« Roman Catholic (RC)
»40« Wesleyan (WES)
»50« Baptist (BAP)
»60« Other Protestant (OP)
»61« Methodist (M)
»62« Primitive Methodist (PM)
»63« United Methodist (UM)
»64« Non Conformist (NC)
»65« Congregationalist (CNG)
»67« Puritan (PU)
»70« Jewish (J)

(Codes 10-60,70 follow the list of religious denominations on attestation forms. The rest were added during recording.)

5.1.3 Three digit codes:

a. Variable name: REASON FOR DISCHARGE
Codes: »100« Demobilization (DEM)
»200« Killed in action (K1A)
»210« Died while in service (DIED)
»300« No longer physically fit (NLPF)
»400« Not likely to become an efficient soldier

(NOTL)
»500« End of engagement (END)
»501« Discharged on re-enlistment to service (RENL)
»502« Under age (UAGE)
»503« Old age (OAGE)
»504« Purchased (PUR)
»509« Not required (NREQ)
»510« Surplus (SURP)
»511« Discharged for civil employment (WORK)
»512« Imprisoned by civil authorities (IMPR)
»513« Misconduct (MISC)
»520« Colonization (COLN)
»530« Commissioned (COMM)
»600« Deserted (DESR)
»666« Missing values
» « Not available

b. Variable name: NEXT OF KIN
Codes:
»100« Mother (M)
»101« Sister (S)
»110« Aunt (A)
»131« Wife (W)
»132« Daughter (Z)
»200« Father (F)
»201« Brother (B)
»210« Uncle (V)
»232« Son (Y)
»666« No relations or friends (X)
»900« Parents (P)
»911« Niece (K)
»912« Nephew (N)
»913« Cousin (C)
»921« Guardian of Children (G)
»929« Friend (E)
»932« Child/Children (D)
»999« Missing values (0,U)
» « Not available

5.2 Coding Places.

c. Variable name: PARISH/TOWN/COUNTY OF BIRTH
PARISH/TOWN/COUNTY OF RESIDENCE
PLACE OF ATTESTATION
PLACE OF RECRUITMENT
ADDRESS OF NEXT OF KIN

All places in Britain and Ireland come under their respective county code, on the basis of the 1911 county boundaries. For example, Smethwick, Birmingham and Warwickshire share the Warwickshire code. Ambiguous categories such as »Preston« or »Richmond« are defined by the coding system as »missing values'; their correct codes could sometime be inferred, on individual basis, from other place-variables of the same observation. Places outside Great Britain are usually grouped on a state level. Therefore, New-York City, California and the USA share one code.

The coding system operates two place-directories. ALLCTY includes some 650 categories which apply to county names alone. The computer first searches this directory, and only after exhausting this source, would activate the ALLPLACE directory which contains over 8000 categories.

5.3 Coding Occupations.

Variable name: OCCUPATION or
SECTOR + OCCUPAT + OCCUHIR

The occupational codes are derived from a classification system created by and described in Michael Anderson et Al. (9). This classification is a further development of Charles Booth's industrial categorization of 1886 (10) corrected and extended by A. Armstrong. (11).

The crux of this classification is that it is industrial rather than occupational^^)

»All persons (including administrative, technical, clerical and ancillary staff) employed in a »unit of industry' are included irrespective of their occupations, in the figures of employment for the industry to which the »unit' is classified.«

The initial breakdown is into nine major industrial sectors, which are further divided into seventy-nine industrial groups:, agriculture 4, mining 4, building 3, manufacturing 31, transport 5, dealing 13, industrial services 2, public service and professional 14, domestic service 3. Each industrial group consists of several occupational headings. For example, the occupational headings »cabinet makers« and »french polishers« are included within the industrial group »furniture« of the manufacturing sector«.

This classification was originally prepared for the analysis of census data. The army's interest in a soldier's occupation differed in many respects from that of the Registrar General. According to their Attestation Forms, recruits were merely asked to specify their »occupation or calling«. No information on the nature of their last job or their last employer was required. Fortunately, many Late Victorian and Edwardian occupations were highly industry specific, and were therefore easily classified. In many other cases the recruits themselves volunteered details on employers or their »unit of industry', and these particulars were recorded on the Attestation Forms as a part of their »occupation and calling«. Complementary information on employers was also abundant in files of wounded or demobilized soldiers (see for example A.F. Z-22) and was added to the occupational entries.

Although each occupational title is allocated a ten figure code, the industrial classification - or SECTOR - occupies only six figures. The first two of the six correspond to the industrial subgroup, while the third figure identifies a particular occupational heading. The latter three figures of SECTOR ensure the allocation of a discrete code to each occupational title.

The four remaining figures of the full code add two additional differentiating concepts to the industrial classification. Each consists of two digits. The first, OCCUPAT, describes the occupational attributes brought to the job market, either as a result of training or as »inborn traits« such as strength or artistic talents. Being a rough estimator as it is, OCCUPAT is based on the assumption that remuneration was closely related to skills. The second, OCCHIR, »summarizes the position [of each occupational title] in the hierarchy of administration, control and task performance«.(13) It provides a much needed differentiation among, for example, foremen, apprentices and labourers. Both OCCUPAT and OCCUHIR are derived from the occupational title itself.

6. Quantitative Assessment of the Sample.

No historical study should expect to obtain a perfectly representative sample of an entity as large and complex as the British ranks of the First World War. However, the First World War archive in Hayes, more than any other contemporary source, draws us nearer such a goal if only for its size and the impressive quality of the data. Yet its quantities may well disguise crucial biases caused by the undocumented loss of many of its files. By creating a representative sample of the archive I merely made sure that if any such biases were introduced to the archive, they would also be accurately reproduced in the sample. Therefore, it would be imprudent to identify the army with either the archive or the sample, without further investigation.

This chapter has, therefore, two aims. The first is to assess the adequacy of the sample as a substitute for the army's ranks. This will be pursued by comparing official statistics with distributions of comparable variables of the sampled data. There are some methodological complications in conducting this comparative exercise, notably, the flaws and ambiguities of the official statistics themselves. It is difficult to reconcile the official figures even on fundamental matters such as the size of the army or the precise number of casualties. However, since it is the validity of the archive that we are interested in, I shall not attempt to rework the official figures. Instead, 1 shall adopt a strictly conservative approach and accept, on face value, those official statistics that were used by other historians in the most recent research; I shall compare them to the sample's distributions and interpret resulting variations as flaws of the archival data. This also means that I shall not attempt to explain the causes for these biases. If flaws they really are, then further research into the history of the archive may be able to describe how they came about.

The second aim of the chapter is to evaluate the differences between the populations in each of the collations and their effect on the full sample. Here too, the method will be comparative, though the statistics will all come from sampled documents. This exercise may provide archivists and researchers alike with additional insight into the representativeness of the collations and their ability to substitute for each other. In both comparisons I shall distinguish among three major groups: »the sample« or »the archive«, i.e. all the sampled observations; »wartime soldiers« including only those discharged after July 1914 and recruited prior to December 1918; and the »wartime recruits« who joined the colours between August 1914 and November 1918. Other subgroups which are not self explanatory, will be described separately, below.

6.1 Army Vs. Sample.

6.1.1 A Comparison of Geographical Spread.

The army never published a detailed breakdown of the geographical distribution of recruiting. Instead, in its official report on wartime recruitment, it grouped the recruits according to their national origin. The similarity between the army and the sample distributions can be seen below in table 8.1, columns 2,3. In the wider context of the total male population in 1914 (column 4), the sample slightly furthers the inherent over representation of English recruits and the under representation of the Welsh in the forces. Unlike the army statistics, it shows a slight under representation of the Scots. The remarkable - though hardly surprising - under representa-

tion of Irish men in the army is fully and accurately captured by the sample.

Table 6.1: National origins of Wartime Recruits.

Country (1)	Army (2)	Sample (3)	Males (4)
England	80.60	84.04	74.21
Wales	5.49	4.12	5.64
Scotland	11.22	9.13	10.45
Ireland	2.69	2.71	9.70
total	100.00	100.00	100.00

Source: P.P. 1921, XX, Cmd. 1193, p.9.

Even more remarkable is the sample's sensitivity to regional variations. Table 6.2 compares a breakdown by counties of the British population of 1911 with the counties of residence of wartime recruits. It also provides, in bold letters, sub-grouping of the distribution into 13 wage regions. (14)

Table 6.2: Spatial Spread of Soldiers and Civilians in Counties and Wage Regions.

County	% in Census	% in Sample
London	11.12	17.40
Surrey	2.07	1.24
Kent	2.56	2.40
Middlesex	2.76	0.74
Essex	3.31	1.34
London & Home Counties	21.80	23.12
Wiltshire	0.70	0.46
Dorset	0.54	0.43
Devon	1.71	1.26
Cornwall	0.80	0.48
Somerset	1.11	0.82
Gloucestershire	1.81	1.75
The Southwest	6.67	5.20
Sussex	1.62	1.32
Hampshire	2.33	1.77
Berks	0.69	0.50
Hertfordshire	0.76	0.78
Buckinghamshire	0.54	0.71

Oxfordshire	0.47	0.52
Northamptonshire	0.85	1.26
Huntingdonshire	0.13	0.11
Bedfordshire	0.48	0.61
Cambridgeshire	0.49	0.41
Suffolk	0.96	1.00
Norfolk	1.22	1.62
The Rural Southeast	10.54	10.61
Monmouthshire & Glamorgan	3.71	3.08
Carmarthen	0.39	0.17
Pembroke	0.22	0.07
South Wales	4.32	3.32
Herefordshire	0.28	0.13
Cardiganshire	0.15	0.06
Breckonshire	0.14	0.04
Radnorshire	0.06	none
Montgom eryshire	0.13	0.24
Flintshire	0.23	0.15
Denbighshire	0.36	0.18
Merionethshire	0.11	0.02
Caernarvon	0.30	0.24
Anglesey	0.12	0.13
Rural Wales & Hereford	1.88	1.19
Shropshire	0.60	0.70
Staffordshire	3.15	3.12
Worcestershire	0.93	0.98
Warwickshire	3.06	3.46
Leicestershire	1.17	1.34
Nottinghamshire	1.48	1.58
Derbyshire	1.66	1.36
Midlands	12.5	12.54
Rutland	0.05	none
Lincolnshire	1.38	1.10
East Riding	1.06	1.22
North Riding	1.03	1.34
Wage Region *7	3.52	3.66
Cheshire	2.36	2.58
Lancashire & Isle of Man	11.66	14.06

West Riding	7.66	8.64
Wage Region #8	21.68	25.28
Cumberland	0.65	0.76
Westmorland	0.16	0.09
Wage Region #9	0.81	0.85
Durham	3.35	3.33
Northumberland	1.71	1.52
Wage Region #10	5.06	4.85
Dumphries	0.18	0.27
Kircudbridge	0.09	none
Wigtonship	0.08	0.04
Peebles	0.04	0.04
Selkirk	0.06	0.15
Roxburgh	0.12	0.09
Berwick	0.07	0.06
South Scotland	0.64	0.65
Ayr	0.66	0.43
Renfrew	0.77	0.59
Dumbarton	0.34	0.22
Lanarkshire	3.54	3.46
Stirling	0.39	0.32
Linlithgow	0.20	0.09
Edinburgh	1.24	1.30
Haddingtonshire	0.11	0.05
Fife	0.66	0.30
Clackmanan	0.08	0.07
Central Scotland	7.99	6.83
Bute	0.04	0.04
Kinross	0.02	0.02
Perth	0.30	0.15
Forfar	0.69	0.45
Kincardine	0.10	0.09
Aberdeenshire	0.76	0.43
Banff	0.15	0.15
Elgin	0.11	0.09
Nairn	0.02	none
Inverness	0.21	0.22
Argyll	0.17	0.11
Ross & Cromarthy	0.19	0.02
Suthernland	0.05	0.09

Caithness	0.08	0.04
Orkney	0.06	none
Shetland	0.07	none
Northern Scotland	3.02	1.90
TOTAL	100.00	100.00

Source: B.R. Mitchell and P. Deane, Abstracts of British Historical Statistics, Population & Vital Statistics Table # 7.

The comparison cannot and is not intended to prove that the soldiering males of 1914-1918 had an identical geographical spread to the whole of the British population in 1911. What the figures so triumphantly tell us is that the sample covers the whole of Britain; it includes 83 out of 89 counties which represent 99.65% of the 1911 population. None of the six counties which constitute the missing third of a percent represented more than one tenth of a percent of the population. In England and Wales they amounted to half that size. Had the archive contained documents from an unrepresentative selection of regiments or record offices, such a widespread representation could not have been possible.

Most counties retained in the sample their original order of magnitude, and often manifested a high degree of similarity. This is easily noticeable in the match between the wage regions. Representation in the sample is, therefore, not only general but also proportional according to the patterns laid down by the census. Spearman's coefficient of rank correlation is a possible measure for this general agreement between wage regions in the sample and in the census. It takes any value between -1 for total disagreement and + 1 where ranking is identical. In our case the coefficient is r = 0.9890. Notable examples for disagreement are Lancashire, London and the Home Counties. The variance which characterizes the latter is largely offset at the level of the wage region. It is probably the combined result of the coding system and the amorphous perception of the administrative boundaries of London among recruiters and recruits alike. For example, a soldier living in Acton, Middlesex, tended to refer to his address as »Acton, London«. Such an address was recorded verbatim from his file and later was coded as »London«. It should be added that independently from the national bias - shown in table 6.1 - table 6.2 reveals that under-represented regions tend to be rural in nature, while the over-represented ones are highly urbanized. Wage region four, South Wales, is a clear example. Like the rest of Wales it has a smaller share in the sample's population than its share in the census. However, the two rural, western counties - Carmarthen and Pembroke - score in the sample a decisively smaller share of their census figures when compared with the more urbanized Monmouthshire and Glamorgan.

Since the comparison here is between civilian pre-war population and a sample of military recruitment, the resulting differences may well be real and represent the actual spread of the wartime army within the civilian population. On the whole, the sample's data on the geographical spread of recruitment is in close accord with official statistics. The fine web of wartime mass recruitment was fully represented in the archive, and hence was also captured by the sample.

6.1.2 Recruitment Over Time.

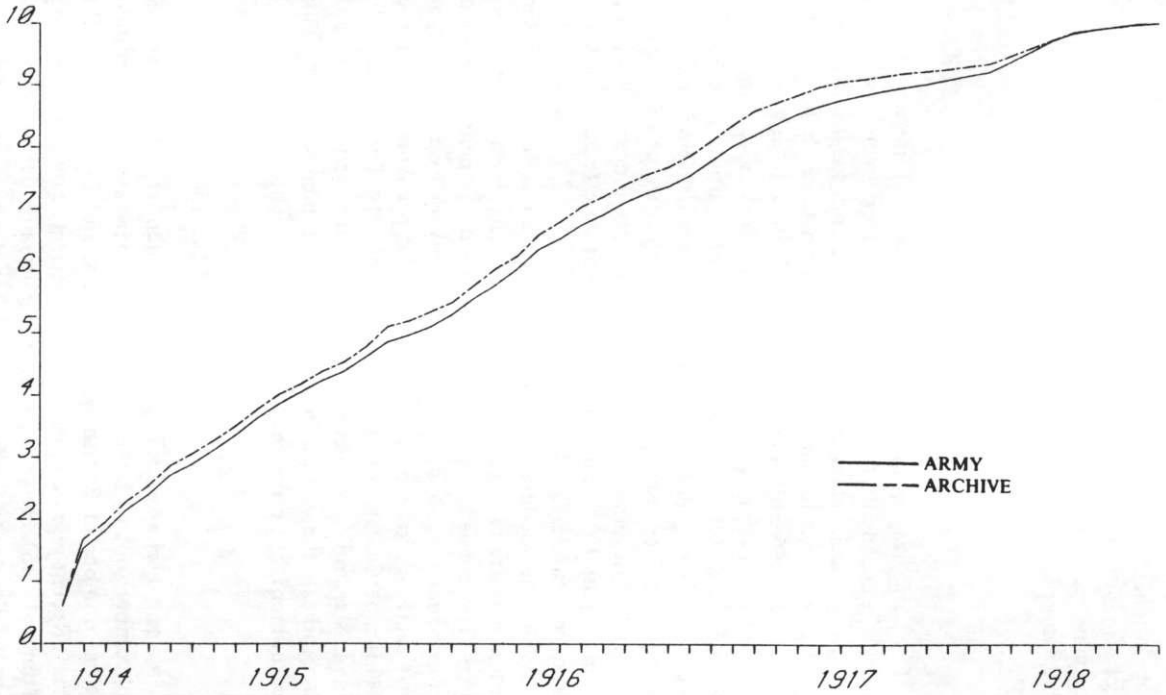
The periodization of recruitment during the First world War has been widely discussed and fairly documented. Several scholars regarded factors as varied as social class, marital status, age, industrial affiliation, fitness or the degree of patriotism, as the explanatory forces behind both voluntary and conscripted enlistment. Since some of these forces were more influential than others during different phases of the war, it is important that the sample - as a proxy for the army - should accurately represent recruiting over time. An over representation of certain periods might distort, for example, the representation of the age structure or social composition of the army. The sample's data on the monthly increments of recruits, between August 1914 and November 1918 are plotted against the army's statistics in figure 6.1 .(15)

Again, the goodness of fit between the time series is clear. Only in five out of the fifty two months do the curves actually move in opposite directions. The persistence of the small gap which opens in September 1914, and reaches a cumulative maximum of only 4% by april 1917, may suggest that there is a true over representation of pre-April 1917 recruits in the archive, particularly of the »great rush« of 1914. This variance, however, is tiny in size and consequence. These results are particularly important since they indicate the unbiased nature of the documents, in portraying not only recruitment, but other related variables as well.

6.1.3 Wartime Casualties.

Among discharged soldiers only the fallen lend themselves to meaningful comparisons with the sample's data. There are several estimates for the actual number of British war casualties, and their share in the wartime army. Recent research has suggested the army's own figures to be »the best estimates (16) However, even the estimates derived from the General Annual Report of the British Army for the war years and for 1920, fluctuate between 9.72% and 12.19% dead out of the whole army. The share of war casualties in the sample amounts to 9.99% of the total wartime sample.

FIGURE 6.1: MONTHLY RATES OF WARTIME RECRUITMENT - AUG 1914 TO NOV 1918
(ARMY Vs SAMPLE)



This share falls well within the expected range of a reliable estimate. This is particularly encouraging because dead soldiers files' are confined to the 'BD' collation and none were added to the archive, via the '14-20' collation, after the Second World War. Indeed, dead soldiers amount to 13.3% of the 'BD', exceeding our upper boundary for a »reliable estimate'. However, the agreement between the archives' data on casualties and other sources goes beyond the similarity in their relative size. In a couple of more rigorous examinations I divided the casualties into age groups (Table 6.3) and compiled a time series of their dates of death (table 6.4) and compared them with official statistics. The p value for the goodness of fit of the age structure is 4.98% and that of the yearly wastage is 29.94%. Together, they confirm the sample's ability to identify sub groups within the army, faithfully represent their unique characteristics and accurately portray their experience over time.

Table 6.3: Age Structure of Wartime Casualties.

Age at Death	% of Sample's Casualties	% of Army's Casualties
16-19	8.91	11.76
20-24	35.61	37.15
25-29	25.28	22.31
30-34	16.85	15.17
35-39	10.32	9.18
40-44	2.23	3.07
45 +	0.80	1.46
total	100.00	100.00

No. of observations: 629.

Source: J. M. Winter, The Great War and the British People p.81. (Based on »The Prudential Life Assurance Company«'s tables for rates of mortality in England & wales.)

Table 6.4: Yearly Rates of Wartime Casualties, August 1914 - September 1919.

Died or Killed in Action Between:	Sample	Army
4. 8.1914 - 31.9.1914	1.0	1.0
1.10.1914- 31.9.1915	14.2	13.9
1.10.1915 - 31.9.1915	22.6	21.8
1.10.1915 - 31.9.1917	25.2	28.4
1.10.1917 - 31.9.1918	27.6	27.6
1.10.1918 - 31.9.1919	9.4	7.3
total	100.0	100.0

No. of observations: 627.
Source: P.P. 1921, XX, Cmd. 1193, pp.62-72.

6.1.4 Religious Denominations.

There are some unique methodological complications in comparing affiliations to religious communities. While it is reasonable to expect no wide disagreements between the distribution of religious denominations among recruits and the share of those denominations among civilians, it should not be surprising to discover an inherent variance between the two groups, which bears no relation to the archive's representativeness. This variance could stem as much from religious dogma as from sociological and demographic peculiarities of religious groups. Some persuasions may advocate conscientious objection and be totally unrepresented in the army; soldiers belonging to religious minority groups may conceal their affiliations, dreading prejudice and stigmatization. Roman catholics, for example, were traditionally over-represented in the prewar ranks while the jews had a higher proportion of their community recruited during the war due to its younger age structure.(17)

There are two further complications in such a comparison. There seems to have been a lesser zeal among recruiters for recording religious denomination. Only 62% of the files contain such information. Secondly, no official army statistics on this matter were published for the war years, nor did the government conduct a religious census since 1851. The only comparable official figures are the returns of the Registrar General on marriages for 1924, which make possible the following breakdown:.

Table 6.5: Religious Affiliations of Soldiers and Civilians.

Religion	% of Sample	% of Civilians
Church of England	72.5	68.8
Presbyterian	8.4	8.3
Roman Catholic	7.5	7.9
Methodist	6.9	7.5
Congregationalist	2.1	2.9
Baptist	1.7	2.4
Jewish	0.9	0.8
Other	0.2	1.4
total	100.0	100.0

Source: The Registrar General, Statistical Review of England and Wales for 1924,11 p.64; The Registrar General For Scotland, 71st Annual Report for 1924, p. lxxxix.

The share of Anglicans in'the military data is higher than their share in the religious marriages, while most of the smaller denominations show higher frequencies. This phenomenon may be partly attributed to the continuous decline in religious marriages among Anglicans, said to be responsible for the only significant decrease in the proportions [of religious marriages] for 1924«, and the tendency of the Roman Catholic, non-conformist and Jewish marriages to increase or to hold to their pre-war proportions. (18) On the whole, however, the match between the distributions is very close indeed. The comparison produced an identical rank order between the two series (r= 1.00) and retained the correct order of magnitude of each group.

Various variables have been examined, all confirming the data's representativeness and their high degree of accuracy in portraying the First World War ranks. It is important to emphasize that these qualities were not confined to general trends and overall averages. The data has proven remarkably sensitive in reproducing religious, geographical and age groups, as well as monitoring short time fluctuations. Nevertheless, some variances from official statistics have also been noticed: a tendency for over-representation of pre-April 1917 recruits, a higher proportion of men residing in urban districts and an underrepresentation of Scotish and Welsh documents. These differences, however, were small and certainly not sufficient to qualify our reliance on the data to any significant extent.

6.2 The 'BD' Vs. the'14-20'.

At present, the '14-20' is clearly the favourite collation. The 'BD' is regarded more as a costly, unescapable burden. Long years of frustration in dealing with an incomplete archive of dusty and crumbling documents took their toll and did not contribute towards the 'BD"s reputation. To add insult to injury, its huge dimensions and apparent chaos hindered a full enquiry into its quality as a historical source. Nor was the suitability of the '14-20' as a substitute for the archive ever tested. The concluding section of this comparison is devoted to such an assessment. Its purpose is to illustrate the consequences of omitting any of the collations from any sample drawn from the archive. In other words, by focussing on the differences between the populations in each collation, I shall point out the nature of the distortions that might have been introduced into our sample had I used only one collation.

6.2.1 Reasons for Discharge.

Some aspects of the differences between the collations were discussed in chapter 2.2 above. The striking disagreement between the »Reasons for Discharge' in table 2.1 suggested profound difference in the military histories of soldiers in each collation. Since table 2.1 is based on a rather limited number of observations, it seems only right to commence our comparison by examining the »Reasons for Discharge' in the full sample. Table 6.6 below fully confirms the initial insight. The basic distinction between the collations remained clear: Soldiers who went through the process of demobilization after the signing of the armistice, constitute the predominant group in the 'BD'. The medically unfit, most of whom must have been released throughout the war, form the majority in the '14-20'. Fatal casualties are virtually excluded from the '14-20' where, on the other hand, the share of soldiers »not likely to become efficient' is twelve times higher. Army authorities in charge of formal releases from service, seem to have dealt less with the 'BD' soldiers. Even among the delinquents, the majority in the 'BD' left the service on their own initiative, i.e. deserted; Most of the '14-20' delinquents were discharged with ignominy or after imprisonment due to a disciplinary process. It would be only logical to expect the circumstances on discharge to be correlated with the history of service.

FIGURE 6.2: MONTHLY RATES OF WARTIME RECRUITMENT - AUG 1914 TO NOV 1918.
(ARMY. SAMPLE, BD', '14-20')

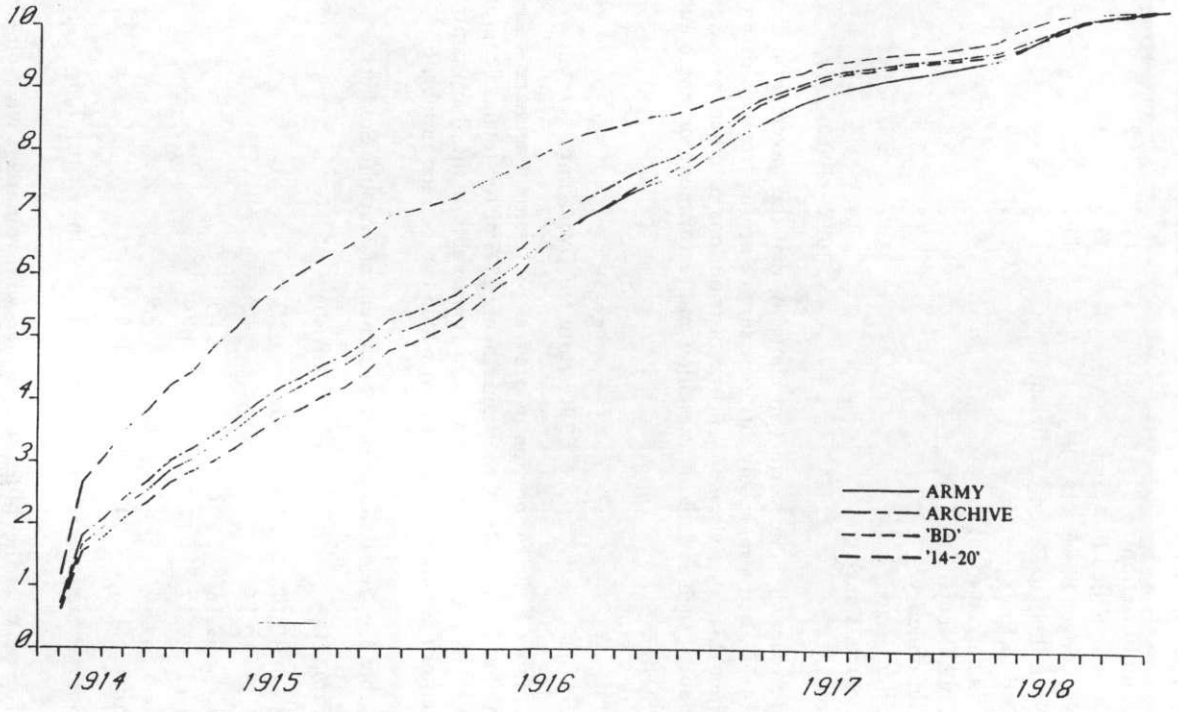


Table 6.6: The 'Reasons for Discharge' of Wartime Soldiers.

Reasons for Discharge	'BD'	'1420'
Demobilization	62.7	23.2
Died or Killed in Action	13.3	0.1
No Longer Physically Fit	11.1	32.8
Not Likely to Become Efficient	1.9	22.6
End of Engagement	3.9	9.9
Discharged to take Civil Occupation	3.8	5.8
Delinquency	2.8	2.1
Under Age	0.2	3.1
Commissioned	0.4	0.4
total	100.0	100.0

However, if I am correct in claiming that the »average soldier« in the '14-20', had a markedly different service profile from his pal in the 'BD', then this must also show in the distributions of other service-related variables, such as dates of attestation and recruitment, or length and place of service.

6.2.2 Recruitment Over Time.

Indeed, the proportion of pre-war professional soldiers is significantly higher in the '14-20' the figure being 21% of the total, as compared with 12% in the 'BD'. But even more remarkable is the sharp contrast in the pattern of wartime recruitment which can be seen in table 6.7.

Table 6.7: Yearly Rates of Recruitment of Wartime Soldiers.

From -to	'BD'	'14-20'	Army
8.14 - 12.14	22.0	36.6	23.7
1.15 - 12.15	25.1	31.5	25.4
1.16 - 12.16	27.6	15.5	24.5
1.17 - 12.17	17.0	9.8	16.5
1.18 - 11.18	8.3	6.6	9.9
total	100.0	100.0	100.0

Source: Statistics of the Military Effort of the British Empire, p.363

The ranks of the '14-20' are swollen with volunteers who joined the army between August 1914 and the introduction of conscription early in 1916. The single largest group in all our wartime distributions is that of the »rush

to the colours« of August and September 1914. However, only among the '14-20' does this group account for a whole quarter of the total. The comparable figures amount to 14.2% in the 'BD' and 15.3% in the official statistics.

Figure 6.2 clearly illustrates the extent of the error that we might have encountered by using only the '14-20' data for describing wartime recruitment. In contrast, the 'BD's curve reasonably follows the patterns set by official statistics. Figure 6.2 also suggests that the two collations are complementary, with the '14-20' more than fully compensating for a slight underrepresentation of the early stages of recruitment in the 'BD'. The differences between the series regarding the post-voluntary period are less dramatic. Thus, slightly less than a half of all conscripts in both collations, attested during the »Derby scheme« of October-December 1915, in acceptable accordance with the official statistics, the figures being 49.1% for the 'BD', 46.5% for the '14-20' and 46.9% for the army. (19)

6.2.3 Discharge Over Time and the Length of Service.

It is worthwhile noting that precisely because the '14-20' is representing particular groups of soldiers, and only amounts to merely one fifth of the data, its gross variations from the true distributions have modest, and usually corrective, effects on the final weighted averages. Further examination of soldiers' military histories confirms this conclusion. Table 6.8 lists the yearly proportions of wartime soldiers discharged from service during the war and on demobilization.

Table 6.8: Yearly Rates of Discharge from Service 1914-1919

Year	'BD'	'14-20'	Total
1914	1.21	13.10	3.84
1915	2.84	18.43	6.29
1916	5.18	15.40	7.44
1917	7.72	12.79	8.84
1918	11.78	12.28	11.84
1919	71.26	28.00	61.68
total	100.00	100.00	100.00

Neither of the collations seem wholly credible. The share of demobilized soldiers is expected to fall around the two-thirds mark, probably nearer 60%. Again, the 'BD' data are nearer the expected target, yet they leave a discernible correction to be desired. This is precisely the role of the '14-20' in the weighted average of 1919. (see total column)

In view of the high rate of human wastage during the war, it might have been expected that an early-enlisting population, such as the '14-20's re-

cruits, would show higher rates of discharge from service in the early years of the war. However, the somewhat constant rate of yearly discharges prior to 1919, and the complete reversal of its pre and post demobilization shares when compared with the 'BD', suggest that other factors must also be at work. Indeed, if we compare the average length of service of individuals in each collation, we find out that the '14-20' also contains a decisive share of soldiers who served for a comparatively short period of time, irrespective of their date of recruitment.

Table 6.9: Length of Service Among Wartime Soldiers

Length of Service	'BD'	'14-20'
Less than 1 month	0.62	4.62
Less than 1 year	13.56	52.46
Between 1-2 years	18.72	15.24
Between 2-3 years	25.06	10.63
Between 3-4 years	24.15	10.83
Between 4-5 years	17.60	10.03
More than 5 years	0.91	0.80
total	100.00	100.00

The share of soldiers who left the army within the very month of their enlistment is seven fold higher in the '14-20'. Just eleven months had elapsed before the majority of the '14-20' soldiers were back in civilian life, while it took almost three years of service to release such a share of 'BD' soldiers.

6.2.4 Place of Service.

The rapid turnover of '14-20' recruits should not be attributed to an extensive participation in battles and to subsequent impairments in health. The data's verdict on this issue is unequivocally negative. Table 6.10 shows a clear negative relationship between service abroad and the likelihood of a wartime recruit being found in the '14-20' collation. Almost two-thirds of the '14-20' served only on British and Irish soil, while an even higher share of the 'BD' served abroad. Note that the incidence of service abroad among pre-war recruits, in both collations, is practically identical.

Table 6.10: Service at Home or Abroad.

Place of Service	'BD'	'14-20'
1. Wartime Recruits		
Britain and Ireland	22.2	62.7
Abroad	78.8	37.3
total	100.0	100.0
2. Pre-war Recruits		
Britain and Ireland	23.9	24.2
Abroad	76.1	75.8
total	100.0	100.0

It may now be argued that, of the two collations it is the 'BD' which reasonably follows the broad characteristics of the wartime army. In contrast, the '14-20' incorporates several overlapping subgroups which are best defined in terms of their exceptional military experience: A strong group of long serving, pre-war professionals, an overwhelming majority of wartime volunteers, a majority of non-combatant, unfit soldiers who left the service not before long and so on. Its user-friendliness has therefore been deceptive. On its own, the '14-20' is clearly unsuitable to serve as a representative sample of the wartime army. However, it would also be wrong to denounce it altogether. First, the '14-20' files were undoubtedly an immanent part of the original First World War archive. Second, where differences occur between the 'BD' and the official statistics, as for example in the time series on recruitment, the '14-20' data play a corrective role in diminishing these variances. Finally, excluding the '14-20' data will deprive the sample of crucial subgroups that are almost exclusively represented in the '14-20'.

To conclude, it has been established that the 'BD' is the backbone of the First World War archive. Its exclusion from any future archive or sample will produce a collection of incoherent remnants, unrepresentative of the wartime ranks. The '14-20' should also be included in any future sample, though in the absence of reliable official statistics, it is difficult to determine its optimal share in the total. Using, as I did, the share of the '14-20' in the remaining archive, produced a sample of remarkable agreement with the official statistics. However, this is only one option and other options should be explored in further research.

7. Conclusions.

The Hayes archive of soldiers' service files is Britain's most comprehensive historical source on individuals who served with the colours during the First World War. The major conclusion of this report is also a crucial contribution towards its vindication as a representative and wealthy source of information. This report also proves that the alleged uselessness of the damaged documents in the archive is unfounded. Another alleged handicap of the archive, the costliness, in terms of time, of extracting a representative sample of meaningful size and significance, cannot be entirely denied. However, this drawback applies mainly to the sample which has been drawn. Future researchers who wish to draw their own sample from the archive, or to complement the existing sample with additional data, could cut the sampling process short by using the knowledge acquired by this research as well as some of its techniques. For many other researchers any sampling from the archive has been rendered unnecessary; to date, three alternative sources provide representative sets of the archive's data. The sampled collation comprises a reliable set of original documents whose process of compilation is documented and available for scrutiny. Since it includes complete files, the choice of variables is entirely left to the researcher. The infinite range of possibilities is illustrated by chapter 2.3, »The Forms«. Some researchers may even be satisfied with the choice of variables and codes described in chapters four and five; in that case, no data recording will be necessary at all. However, it is more likely that future scholars, encouraged by the quality of the sample, will use both the sample and the full archive in order to create specific databases that will suit their own interests.

The issue of representativeness is of the utmost importance. At this stage, I am still unable to explain fully the mechanism that has brought about this representative survival' of the documents. Neither have I set my mind to solving such a mystery. Instead, I preferred to enquire into the apparent influences of past events on the archive's population. Of the two collations the 'BD' is the most reliable. Many historians who are used to the harshness of analyzing incomplete data, may even regard it as a decent substitute for the whole of the army's ranks. The '14-20' is clearly unsuited to serve in such a role. Yet, together, they preform as complementary elements to produce an accurate picture of the army both in general and in detail.

Notes and References.

* I wish to thank the heads of the Ministry of Defence departments CS(R) and CS(R)II for their cooperation. Roderick Floud, Avner Offer, Annabel Gregory and Bernard Harris kindly read earlier versions of this paper and I wish to thank them for their helpful comments.

(1) T. Barker and M. Drake (eds.), Population and Society in Britain, 1850-1980, (London 1982) p.11.

(2) The Report of the Committee on Departmental Records (Cmd. 9163) RR 1954, cited in A.A.H. Knightbridge, »Particular Instance Papers«, Public Record Office, RAD Occasional Papers, No. 8, (London 1984).

(3) R. Floud, K.W.W Wachter, A. Gregory, »The Physical State of the British Working Class, 1870-1914: Evidence from Army Recruits«. National Bureau for Economic Research, Working Paper Series No. 1161, (Cambridge MA. 1985) p. 18.

(4) Annual Report of the Inspector General of Recruiting for the year 1900, (Cd. 519) RR 1901, IX p. 37; Report of the Royal Commission on the War in South Africa, (cd. 1789) RR 1904, XL, pp.41-45, 65; and A.R. Skelley, The Victorian Army at Home, (London 1977) pp. 298, 301.

(5) General Annual Reports of the British Army 1913 - 1919, (Cmd. 1193) RP 1921, XX, p.6.

(6) See for example J.G. Adami, »The Physical Census«, in Transactions of the Medical Society of London Vol.XVII, 1919; Report upon the Physical Examination of Men of Military Age by National Service Medical boards, (Cmd. 504) RR 1919, XXVI, p.22; For later works, J. Oddy, »The Health of the People« in T. Barker and M. Drake (eds.) Population and Society in Britain 1850-1980, (London 1982) p.129 and a critical assessment by J.M. Winter »Military Fitness and Civilian Health in Britain During the First World War« Journal of Contemporary History Vol.15,1980, passim.

(7) J.M. Winter, »The Preparation of an Archive of Medical Disablement Records of Pensioners of the First World War« in Public Record Office, RAD Occasional Papers No.8, (London 1984).

(8) D. Lamm, »The Use of the Portable EPSON PX-8 for Data Collection«, Computing and History Today 2, 1987, p. 18.

(9) M. Anderson et AL, National Sample from the 1851 Census of Great Britain, Background Paper VII: Data Set Documentation (Preliminary Version), (Edinburgh 1977).

(10) C. Booth, Occupations of the People: England, Scotland, Ireland 1841-1881, (London 1886).

(11) A. Armstrong, »The Use of Information about Occupation« in E.A. Wrigley, (cd.) Nineteenth Century Society: Essays in the Use of Quantitative Methods for the Study of Social Data, (London 1972).

(12) M. Anderson et Al. op. cit. p. D.7.e.

(13) M. Anderson et Al. op. cit. p. D.10.

(14) The grouping into »wage regions« is that of E.H. Hunt, Regional Wage Variations in Britain 185Q-1914,(Oxford 1973).

(15) Statistics of the Military Effort of the British Empire during the Great War, (HMSO 1922) p.363.

(16) J.M. Winter, »Some Aspects of the Demographic Consequences of the First World War in Britain« Population Studies, XXX 1976, p.541.

(17) On Irish soldiers see: A.R. Skelly, op. cit. pp.166, 284-287; on Jewish soldiers: B.A. Kosmin, S.Waterman, N.Grizzard, »The Jewish Dead in the Great War as an Indicator for the Location, Size and Social Structure of Anglo-Jewry in 1914« Immigrants & Minorities, Vol. 5,2 1986, p.183.

(18) The Registrar General's Statistical Review of England and Wales for 1924 part II - civil, (HMSO 1925) pp.134-135.

(19) Public Record Office, NATS1/400; Statistics of the Military Effort, p.363.